V. Association analysis II

- Genomewide association studies
- Handling of large data
- Multiple testing
- Graphical methods
- Meta-analysis
- Case studies

Genomewide association studies

- Abbreviated as GWAS
- Typically involves 10k~1M SNPs
- It is unbiased in the sense that SNPs are given the same a priori probability of association.
- It has impact on other mapping methods such as admixture mapping and the *de facto* method of association analysis.
- The have now been used most of the common diseases and traits and leading to a flood of publications. At the meantime, a range of issues of appropriate use genome data are yet to put into practice.

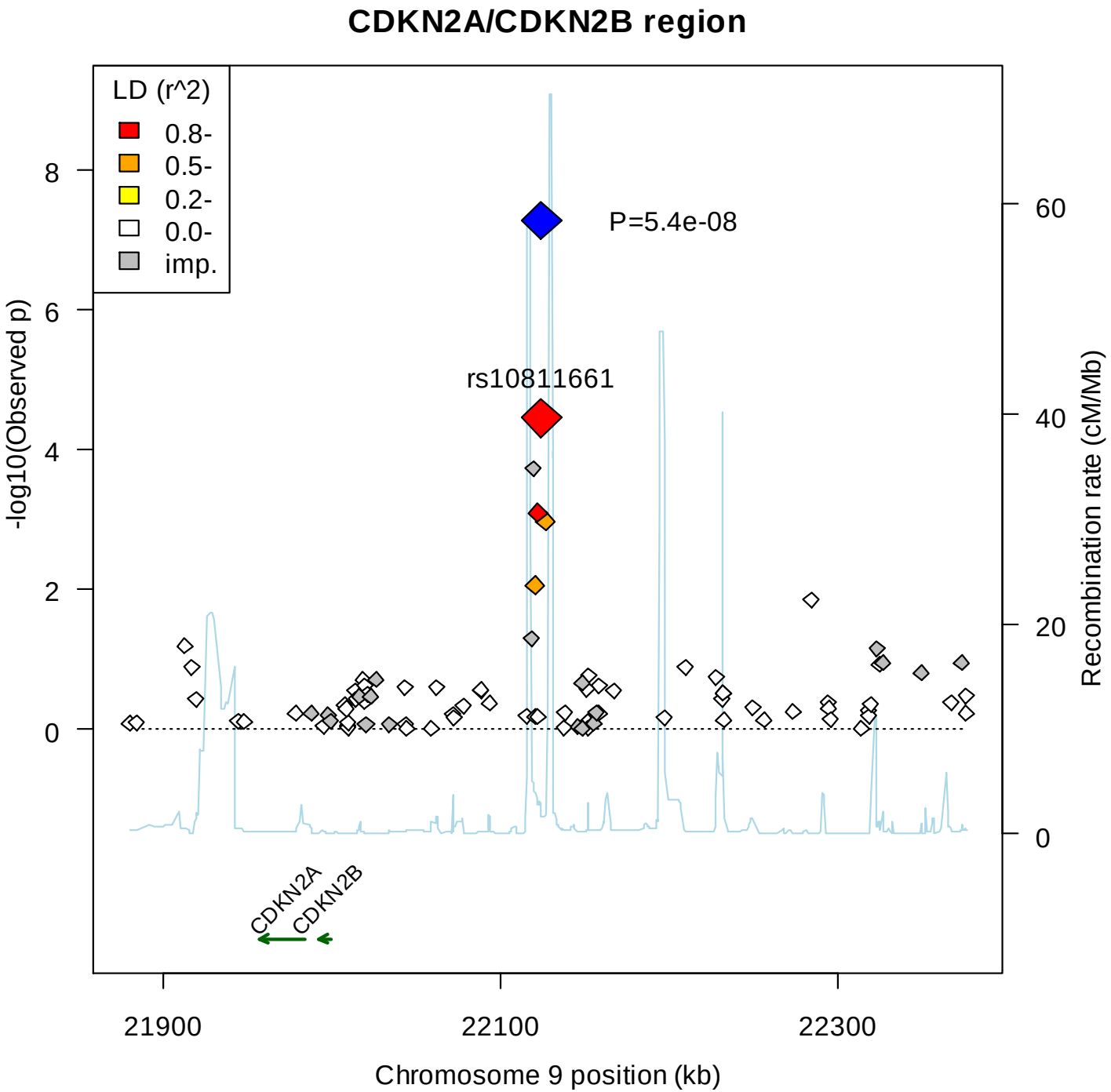
The HapMap project

- The project developed a haplotype map of the human genome, which describes the common patterns of human DNA sequence variation. It is a key resource for researchers to use to find genes affecting health, disease, and responses to drugs and environmental factors. The information produced by the project is freely available.
- The project is a collaboration among scientists in Japan, the U.K., Canada, China, Nigeria, and the U.S. The project officially started on October 27-29, 2002 (http://genome.gov/10005336)
- It has been used to examine LD structure and Infer untyped genotypes (imputation)
- A recent initiative is to expand the project from 270 individuals to 1000.

Manolio et al. (2008) *J Clin Invest* 118:1590-1605

LD information from HapMap

```
%let url=http://www.hapmap.org/downloads/ld_data/latest;
%let file=ld_chr22_CEU.txt.gz;
filename ldin url "&url/&file" recfm=S;
filename ldout "&file";
data _null_;
        infile ldin;
        file ldout recfm=F;
        input;
        put _infile_;
run;
filename in2 pipe "gzip -dcq &file";
data x.ld;
        infile in2 dlm=' ' lrecl=200 firstobs=2 n=2000;
        format pop $3. rs1 $15. rs2 $15.;
        input pos1 pos2 pop rs1 rs2 dprime r2 lod fbin;
run;
```


CDKN2A/CDKN2B region
LD (r^2)
0.8-
0.5-
0.2-
0.0-
imp.
P=5.4e-08
rs10811661
CDKN2A
CDKN2B
-log10(Observed p)
Recombination rate (cM/Mb)
Chromosome 9 position (kb)
21900
22100
22300
0
2
4
6
8
0
20
40
60

Large data

- Database management systems might be required, as may be already in place in some epidemiological cohorts
- If a SNP is stored as 1 byte, for a 1M GeneChip and 10,000 individuals one needs 10GB storage. One may not be able to click to examine the raw data
- We can employ specialized programs for standard analysis
- General systems can be viable for analysis of a range of genetic and nongenetic factors

The wide format

ID	Age	Sex	BMI	SNP1
WTCCC139236	60	2	34.13	A/C
WTCCC139239	60	2	31.00	CC
WTCCC139240	50	2	30.03	AA
…				

The long format

ID	rsn	Pos	age	sex	BMI	Add	Dom	Rec
1	snp1	1231	60	1	30	0	0	0
1	snp2	7891	60	1	30	1	1	0
1	snp3	12321	60	1	30	2	1	1
2	snp1	12331	30	2	35	2	1	1
2	snp2	15312	30	2	35	1	1	0
2	snp3	22312	30	2	35	0	0	0
…								

The transposed format

ID	WTCCC139236	WTCCC139239	WTCCC139240	**…**
Age	60	60	50	
Sex	2	2	2	
BMI	34.13	31.00	30.03	
SNP1	A/C	CC	AA	
SNP2				
…				

The imputed format

chr	rsn	Pos	A1	A2	Genotypes of ID1			...		
1	snp1	1245	A	C	0.00	0.12	0.88			
1	snp2	1260	C	T	0.80	0.20	0.00			
...										

Conversions

- They may be required in a variety of scenarios
- They can be done to a different degrees by
  - PROC TRANSPOSE/IML (**SAS**)
  - FLIP (**SPSS**)
  - reshape (**Stata**, **R**)
  - awk/gawk/nawk, ruby, Perl, TCL/TK, Java
  - **LINKAGE**, MERLIN, HaploView, GTOOL, PLINK, BC/SNPmax?
  - C/C++/C#, Fortran
- Because statistical packages assign data items in specific formats, they tend to be less efficient than customized routines, e.g. C++.
- A good compromise can be achieved with awk, for which we have a set of routines to be used as exercise.

Data extraction

- The master file is in long format containing all individuals and all SNPs. Now the following code can be used to extract a subset of individuals with their SNPs on a specific chromosome i.

- proc sql;
- create table genotype as select * from long&i
- where id in (select id from id) &
- rsn in (select rsn from map&i) order by id, rsn;
- quit;

- We can use keep in data step but a simpler grep for ASCII files in transposed format.

Linux clusters

- Linux has utilities such as mpirun.
- SAS/Connect can use heterogeneous systems
- It is most useful to use ssh:
  - export GWA=/data/genetics/gwas/3-4-8
  - export RUN=$GWA/id.sh
  - ssh -f c16 "cd $GWA;bash $RUN fatpct 1 12"
  - ssh -f c15 "cd $GWA;bash $RUN fatpct1 1 12"
  - …
  - ssh -f c09 "cd $GWA;bash $RUN fatpct 13 22"
  - ssh -f c08 "cd $GWA;bash $RUN fatpct1 13 22"
  - ssh c16 "ps x"
  - ssh c15 "ps x"
  - …
  - ssh c09 "ps x"
  - ssh c08 "ps x"

Allele-coding

The bottleneck has been allele-coding, but the inclusion of map information would do away with it.

Allelic coding when the minor allele A is coded as B by alphabetical order

Correct Model	Genotype coding			Coded Model	Genotype coding			Change direction of effect
	A/A	A/B	B/B		A/A	A/B	B/B	
Additive	2	1	0	Additive	0	1	2	Yes
Dominant	1	1	0	Recessive	0	0	1	Yes
Recessive	1	0	0	Dominant	0	1	1	Yes

Multiple testing

- The false discovery rate measures the proportion of false positives among all SNPs called significant:

$$FDR = \left[\frac{F}{S} \mid S > 0\right] p(S > 0) \qquad \text{or} \qquad pFDR = E\left[\frac{F}{S} \mid S > 0\right]$$

  where S is the total number of tests called significant but only F of which are true. The p-value is a measure of significance in terms of false positive rate (Type I error rate). The *q*-value is an FDR measure of significance associate with each SNP.

- Since $q - value(p_i) = \min_{t \geq p_i} pFDR(t) \forall t \in [0,1]$ and for a large number of SNPs, we can use FDR for the estimate of *q*-value. The parameter λ is chosen to tune for the proportion of truly null π_0 to obtain the FDR estimate or approximately *q*-value.

- The problem recently has largely been gotten around using a threshold of genomewide significance plus replication.

The Prior Probability of Linkage and FDR

- Morton (1955) set the prior prbability of linkage $\pi \approx 1/20$, for a given type I error rate α, and power W, we have

$$FDR = \frac{\alpha(1-\pi)}{\alpha(1-\pi)+\pi(1-\beta)} = \frac{10\alpha}{19\alpha+(1-\beta)}$$

- Smith (1953) assumed $\pi = 1/24, \alpha = 0.05, 1-\beta = 1$,

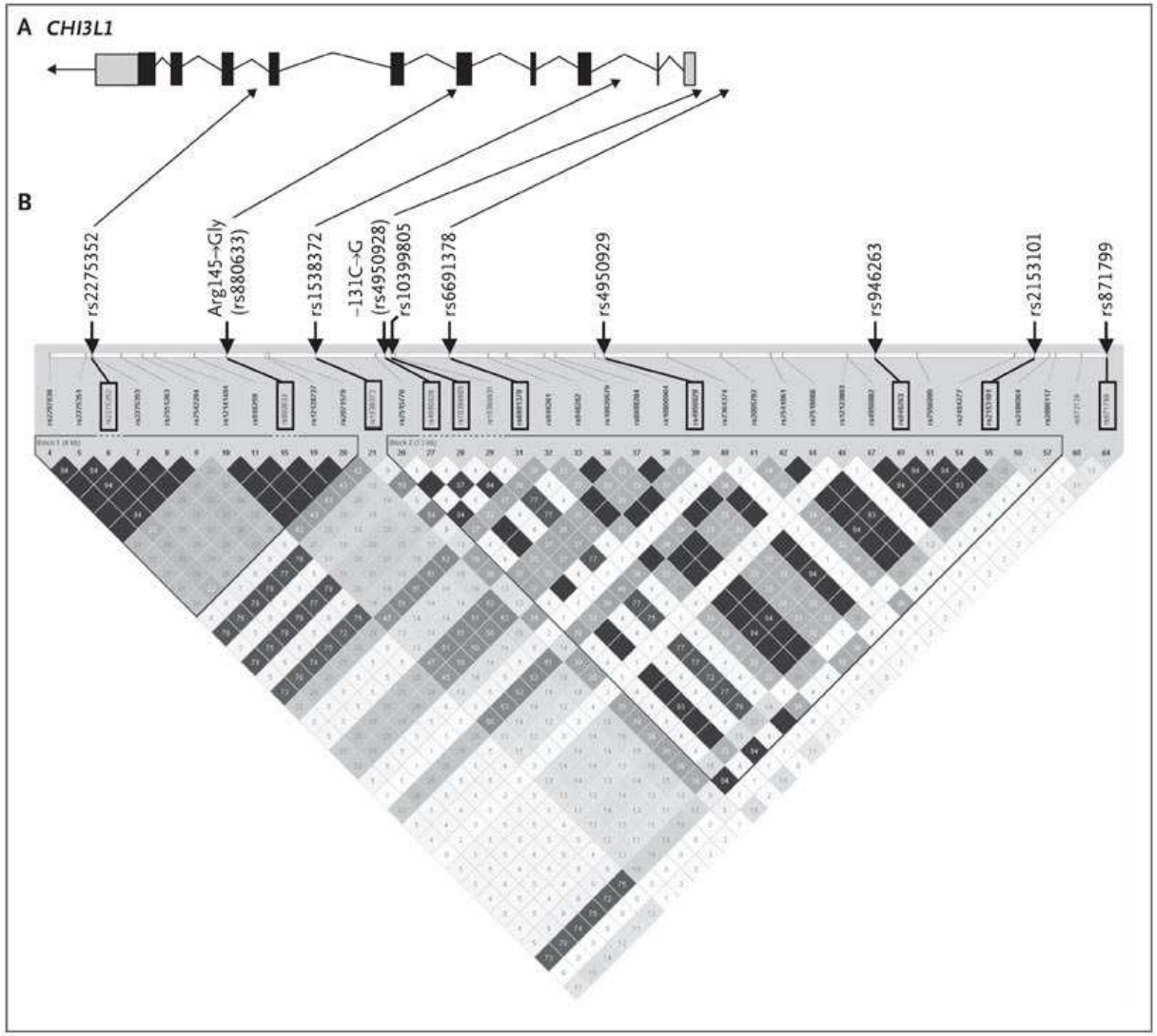
$$FDR = (23/20)(23/20+1) = 0.535$$

- The relationship between type I error α and power $1-\beta$. Since $\alpha = \int_C f_0(x)dx, 1-\beta = \int_C f_1(x)dx$, we have $\alpha/(1-\beta) = \int_C f_0(x)/\int_C f_1(x)dx = \int_C (f_0(x)/f_1(x))f_1(x)dx/\int_C f_1(x)dx = E_1(f_0(x)/f_1(x)|x \in C)$. So if C is region $f_0/f_1 \leq 1/A$, then $\alpha/(1-\beta) \leq 1/A$. For the SPRT, we set $\alpha/(1-\beta) = 1/A$ or $A = (1-\beta)/\alpha$ and $B = \beta/(1-\alpha)$

- These are linked with FPRP of Wachholder et al. (2004) and BFDP of Wakefield (2007)

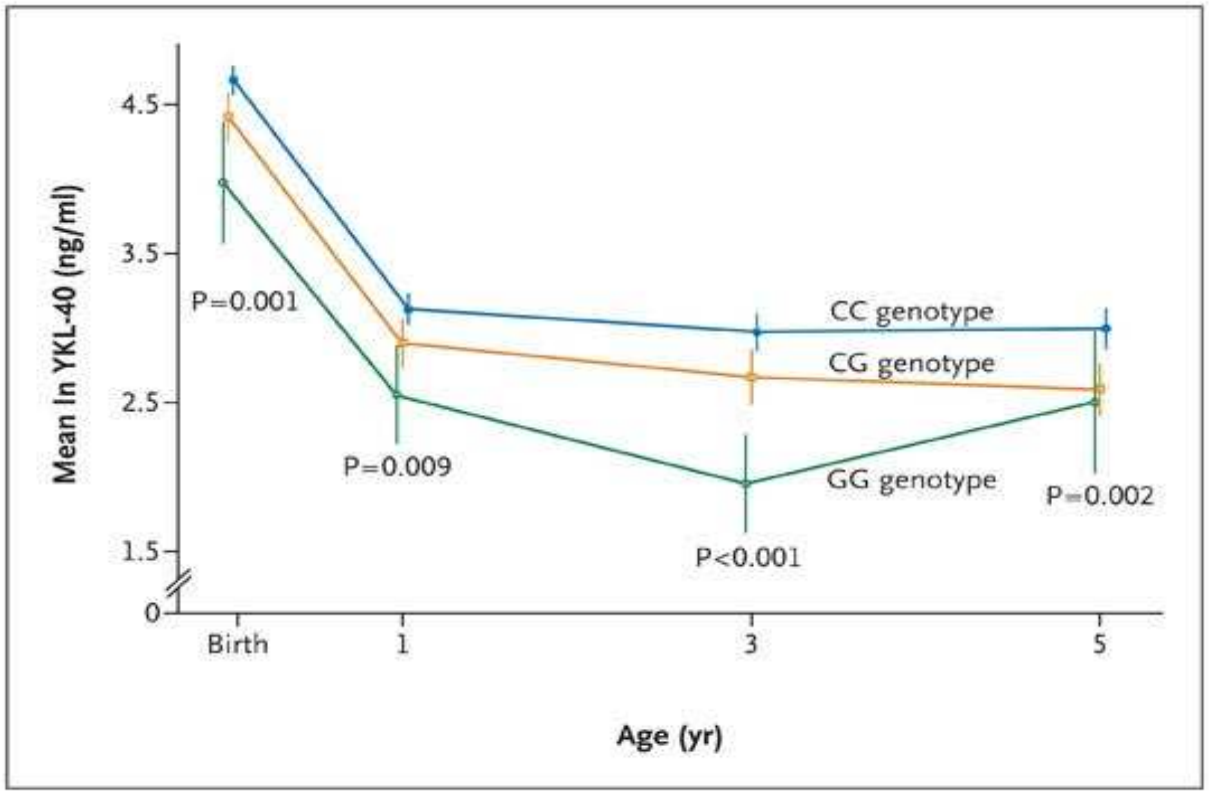
Graphical methods

- Phenotypic data
  - Individual data, e.g., two-way plot, conditional plot
  - Summary statistics
  - Specific features, e.g., pedigree diagram
- Genotypic data
  - Genome level, regional level, functional level
- Genotype-phenotype correlation
  - Q-Q plot
  - Manhattan plot
  - Regional plot
  - Forest plot
  - Receiver-operating-characteristic (ROC) curve



Single-Nucleotide polymorphisms (SNPs) in CHI3L1 and its upstream region on chromosome 1q32.1

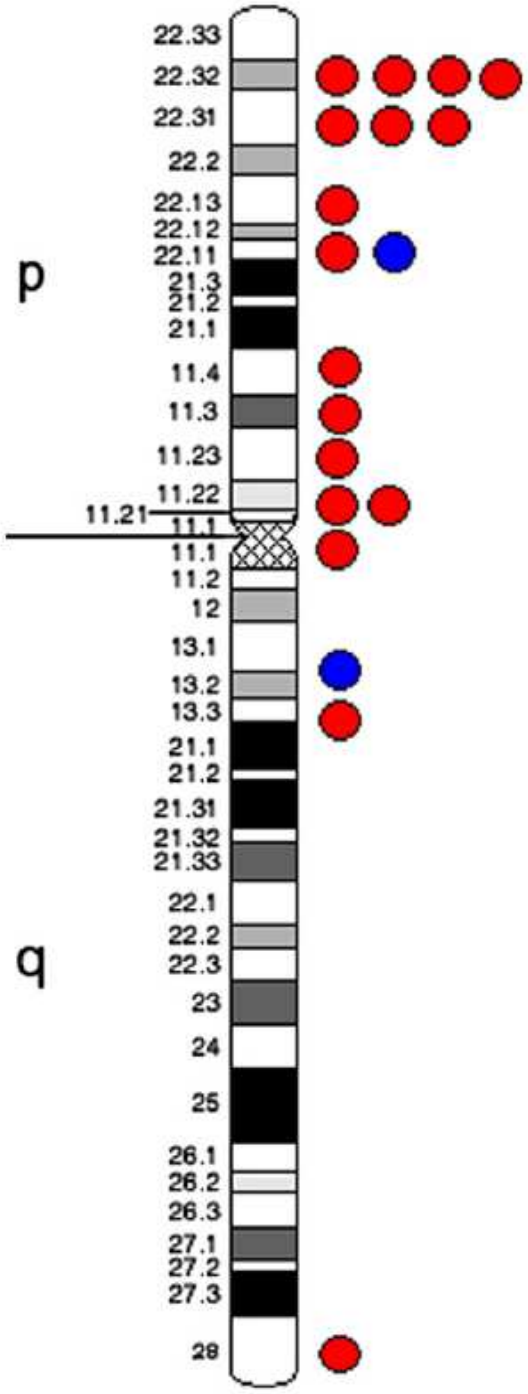
Ober et al. NEJM 2008

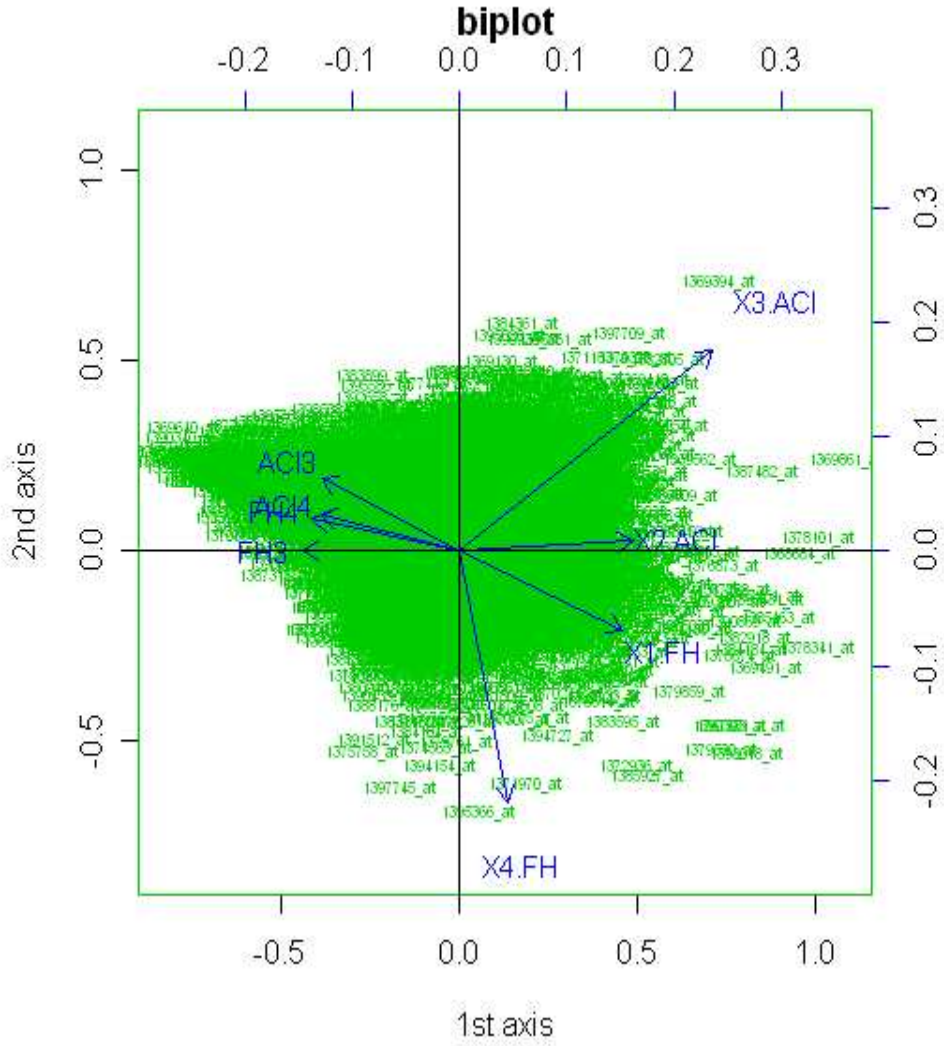


Mean serum YKL-40 levels in Asthma

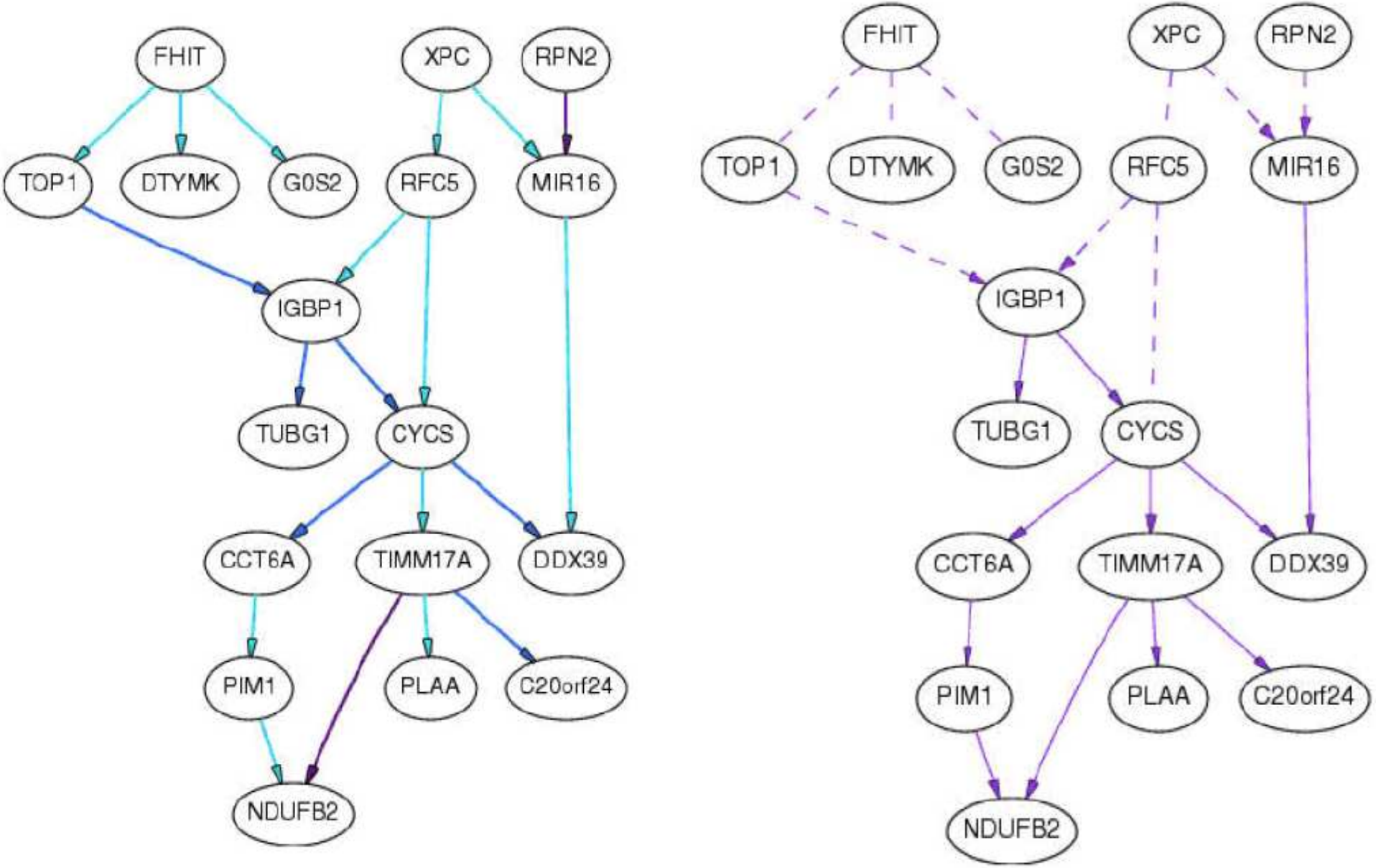
Ober et al. *NEJM* 2008

	CC Genotype			CG Genotype			GG Genotype			*P*-value
	N	Mean	SE	N	Mean	SE	N	Mean	SE	
Cord Blood	82	4.66	0.048	39	4.41	0.082	4	3.98	0.207	0.0010
Year 1	82	3.12	0.054	39	2.90	0.081	4	2.55	0.166	0.0089
Year 3	82	2.97	0.063	39	2.67	0.092	4	1.95	0.169	0.00025
Year 5	71	2.99	0.067	30	2.59	0.090	4	2.50	0.240	0.0016

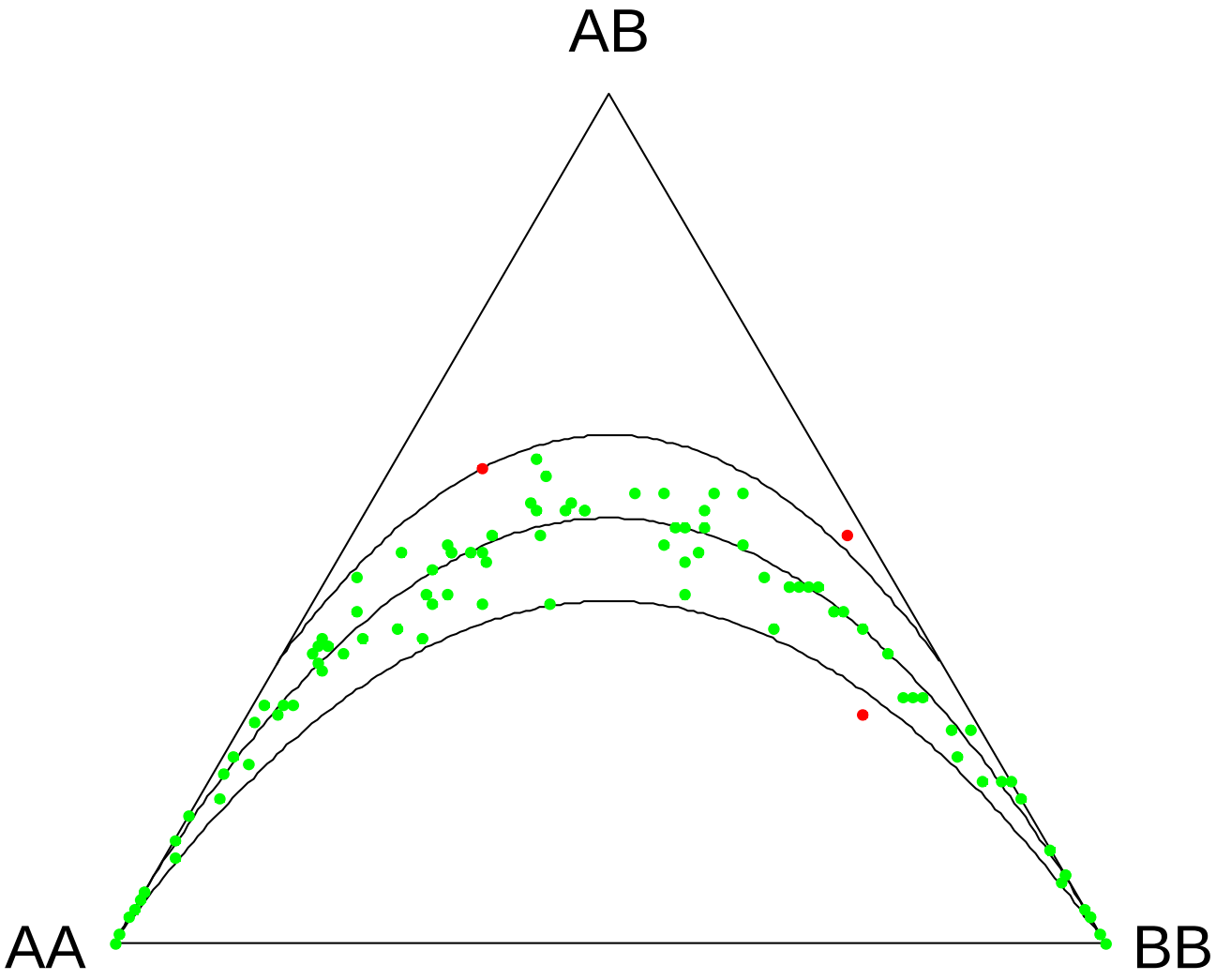




Tan et al. *Genomics* 2008 (and unpublished)

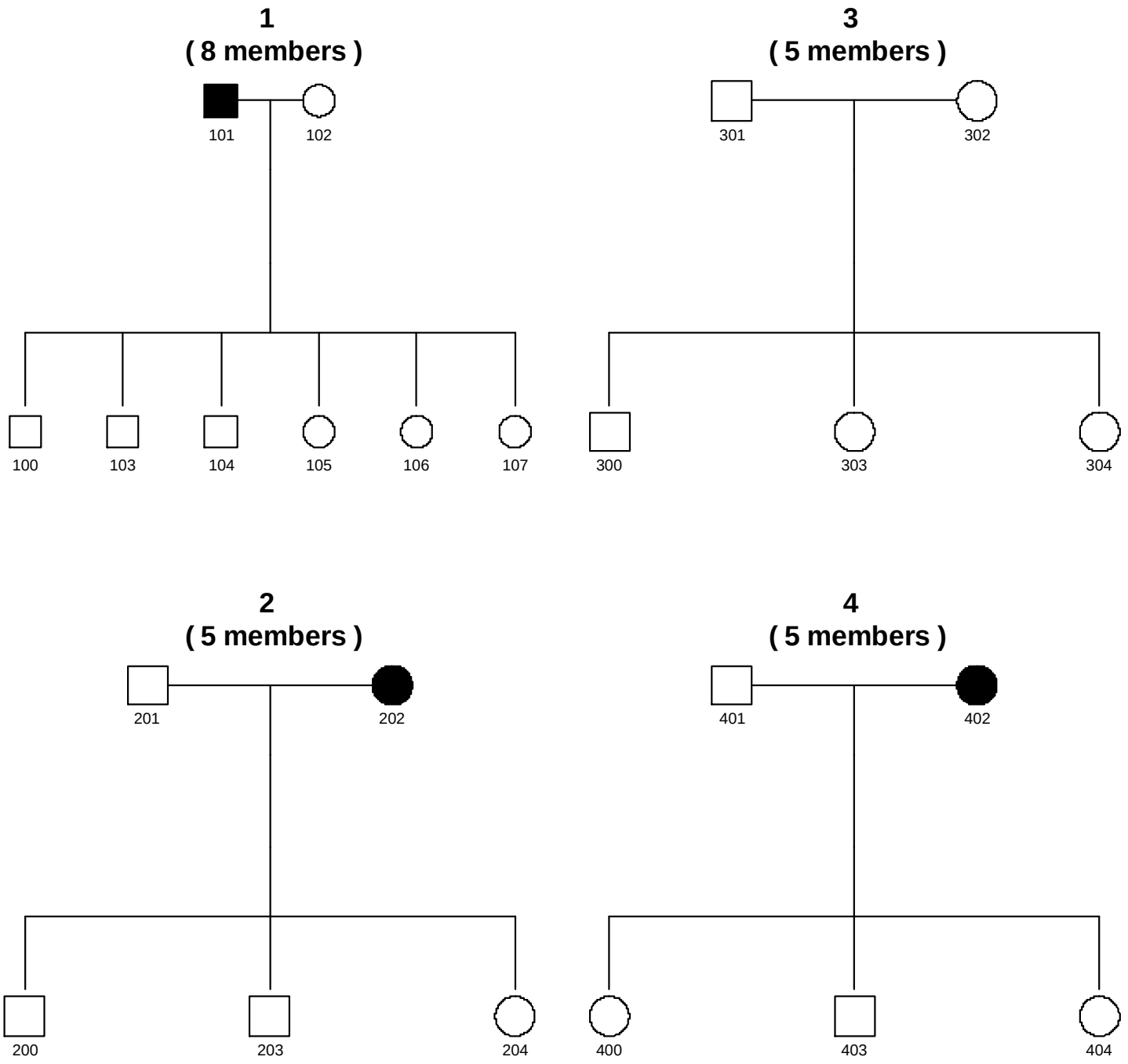


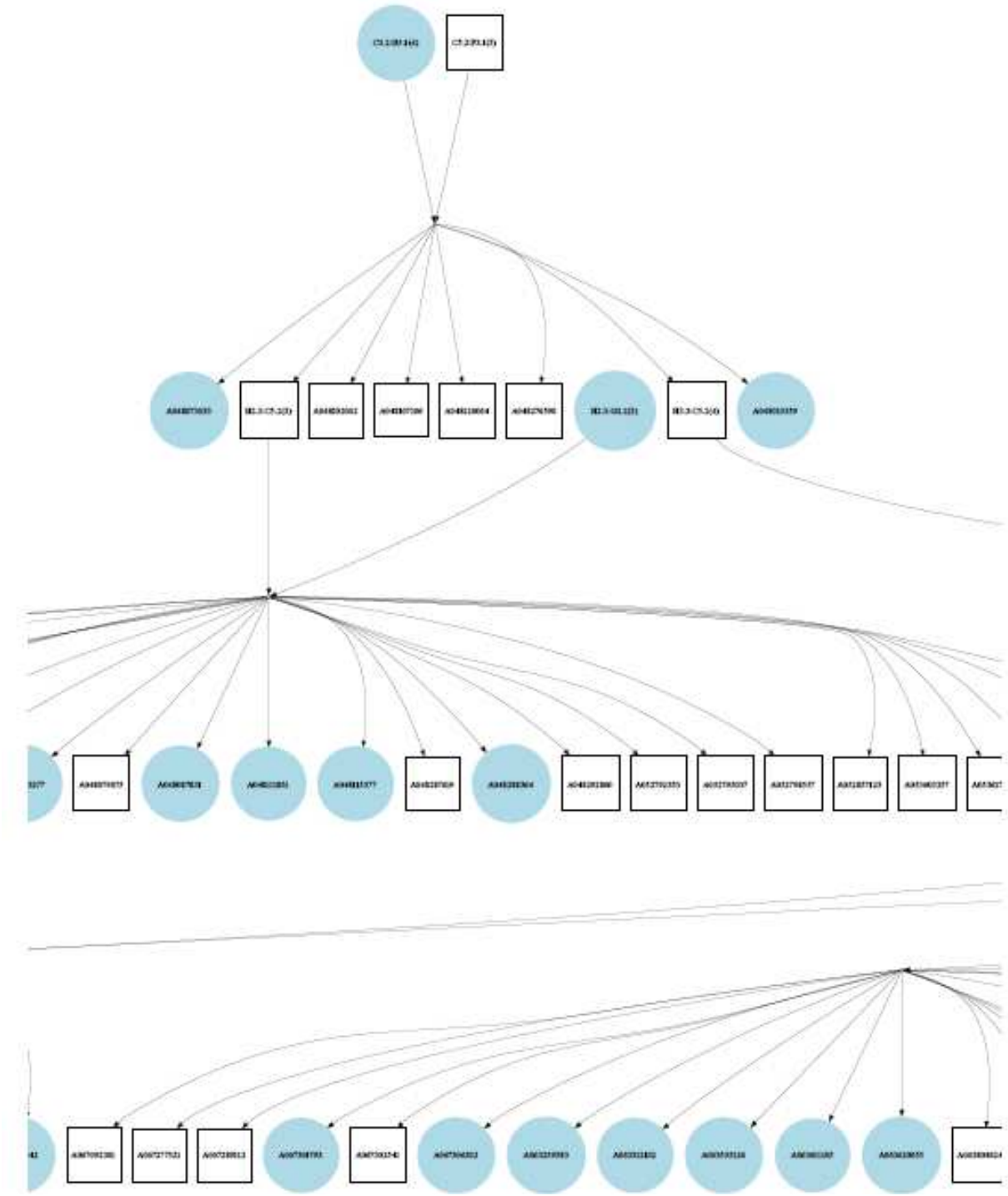
Zhao et al. *BMC Proc* 2007



Ternary plot showing distributions of 100 markers for 100 SNPs

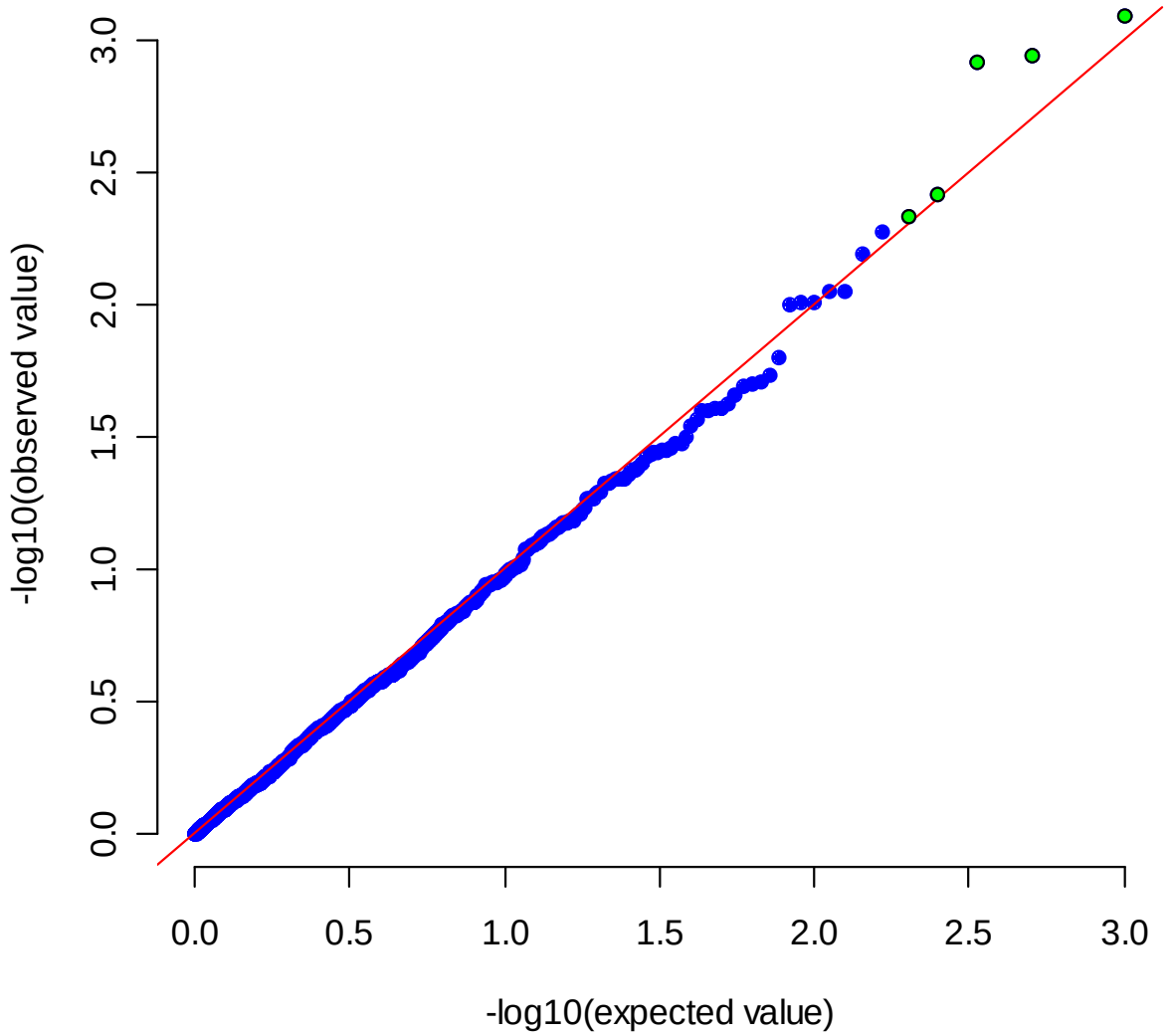
Graffelman & Morales-Camarena *Hum Hered* 2008


1
(8 members)
101
102
100
103
104
105
106
107
3
(5 members)
301
302
300
303
304
2
(5 members)
201
202
200
203
204
4
(5 members)
401
402
400
403
404

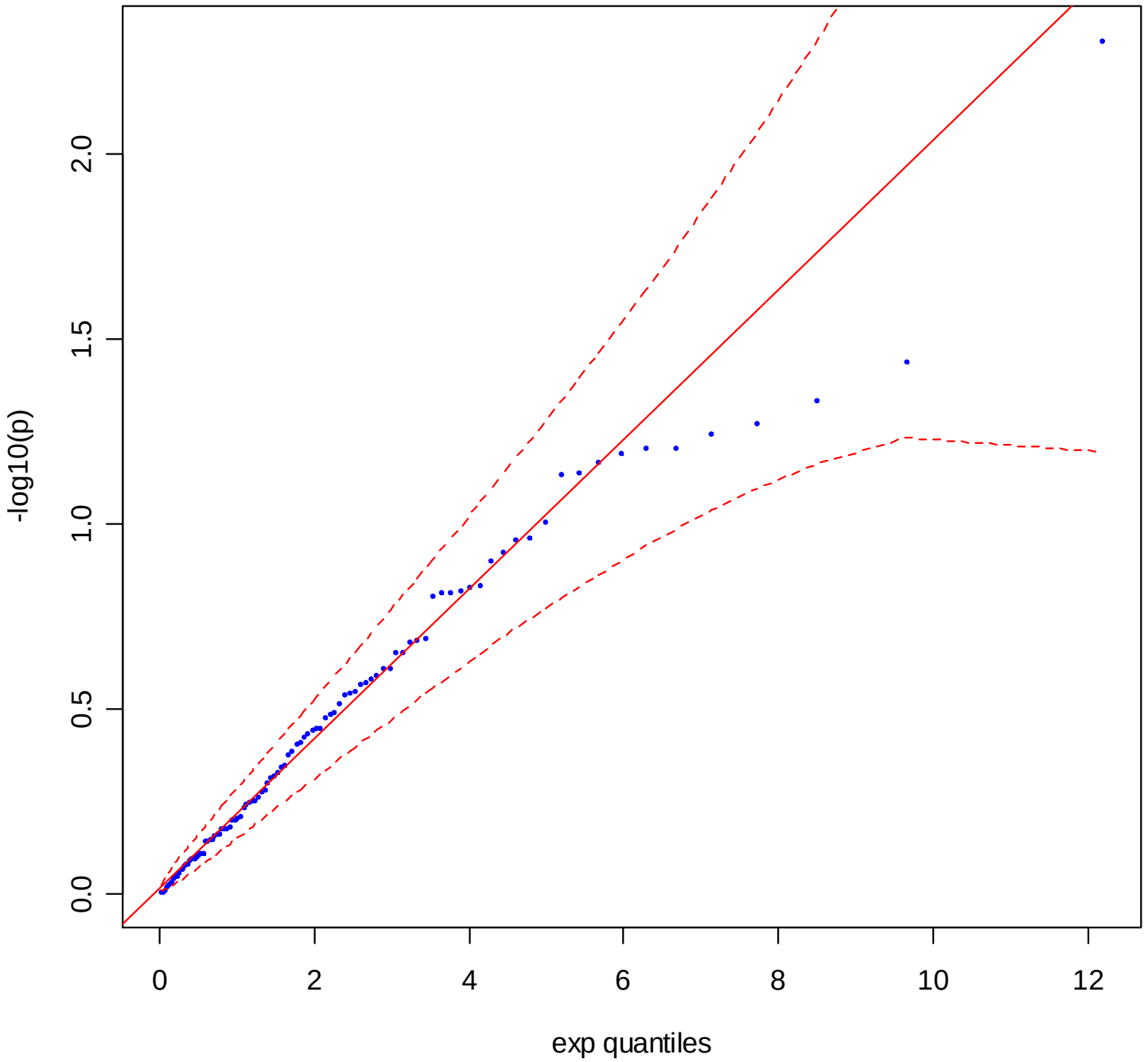


Part of the mouse pedigree from Richard Mott

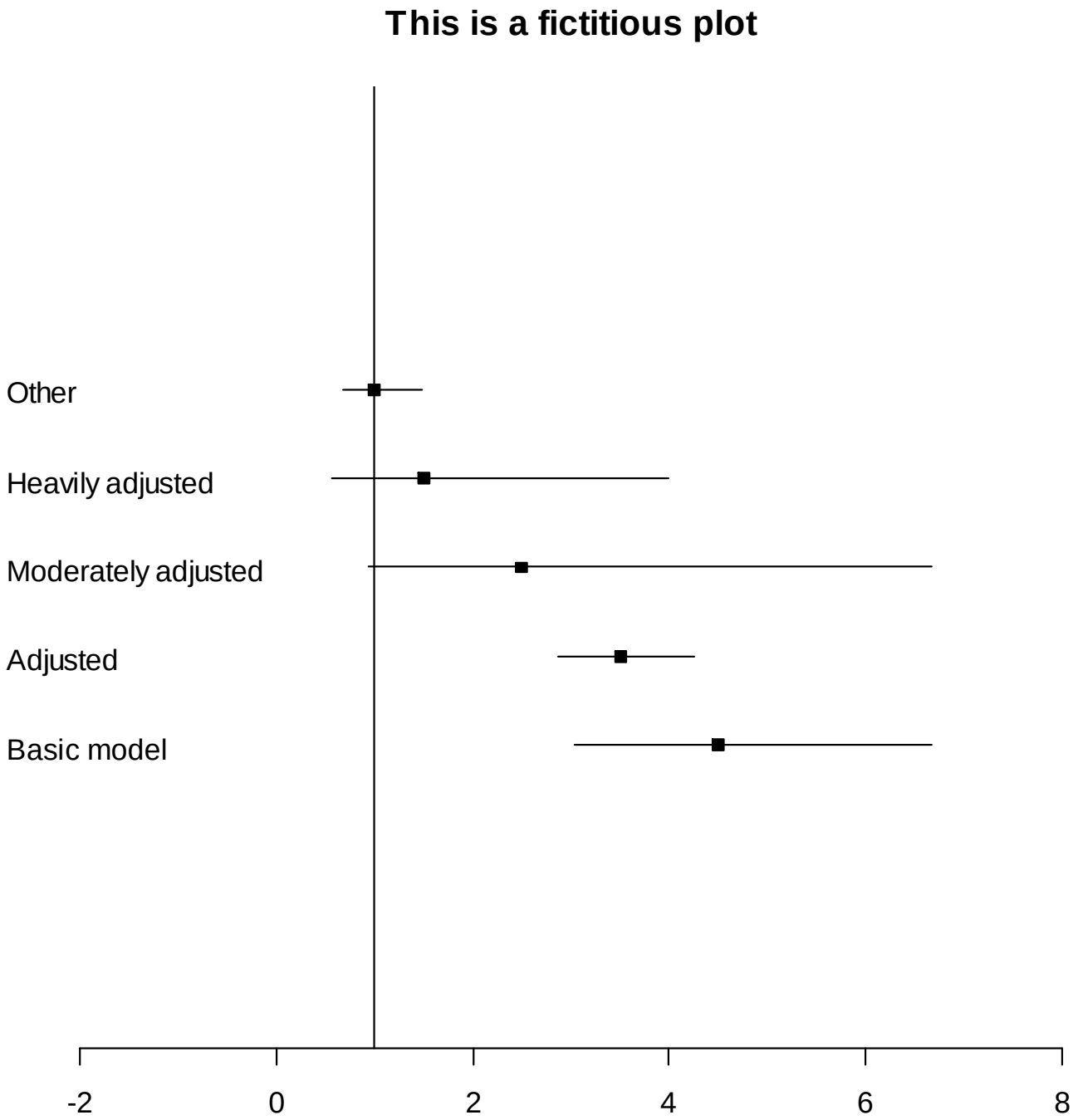
Similar functionality exists in Rgraphviz library but ideally it can also accept .dot file directly



But this is unlike qq.plot, qqmath, the former uses robust statistics

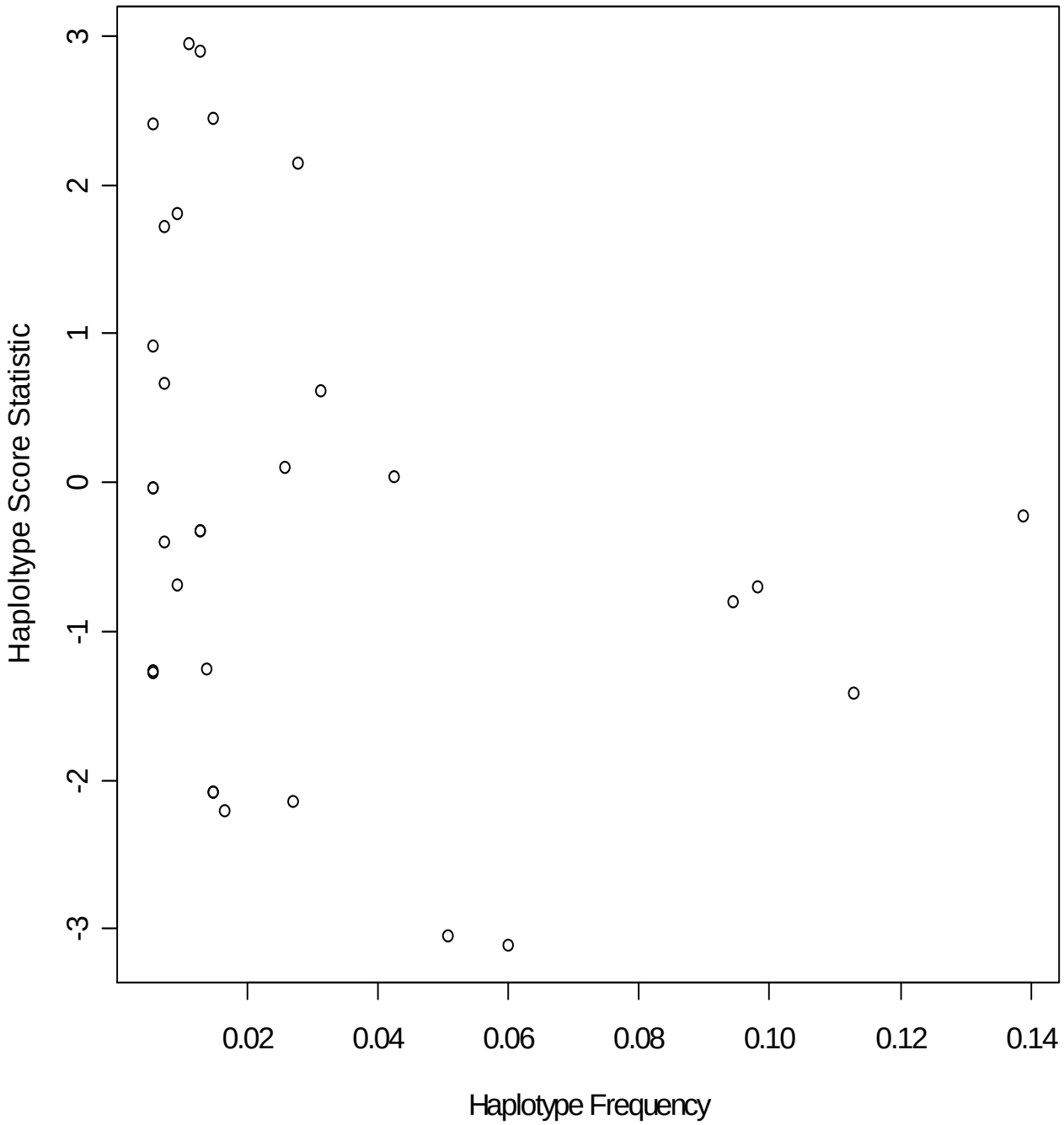


A 95%CI is added, based generally on the order statistics

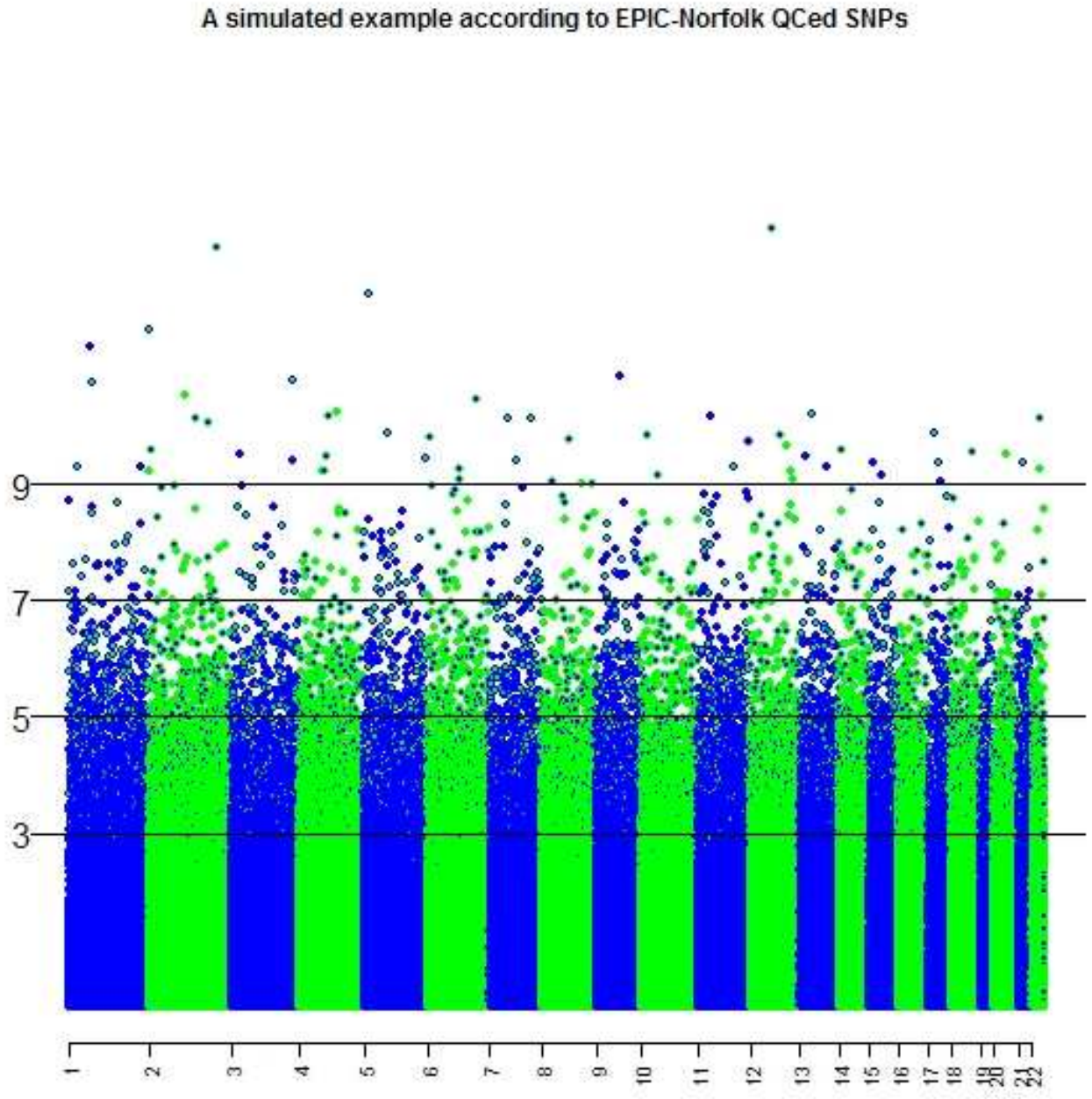


A way of effect-size visualisation

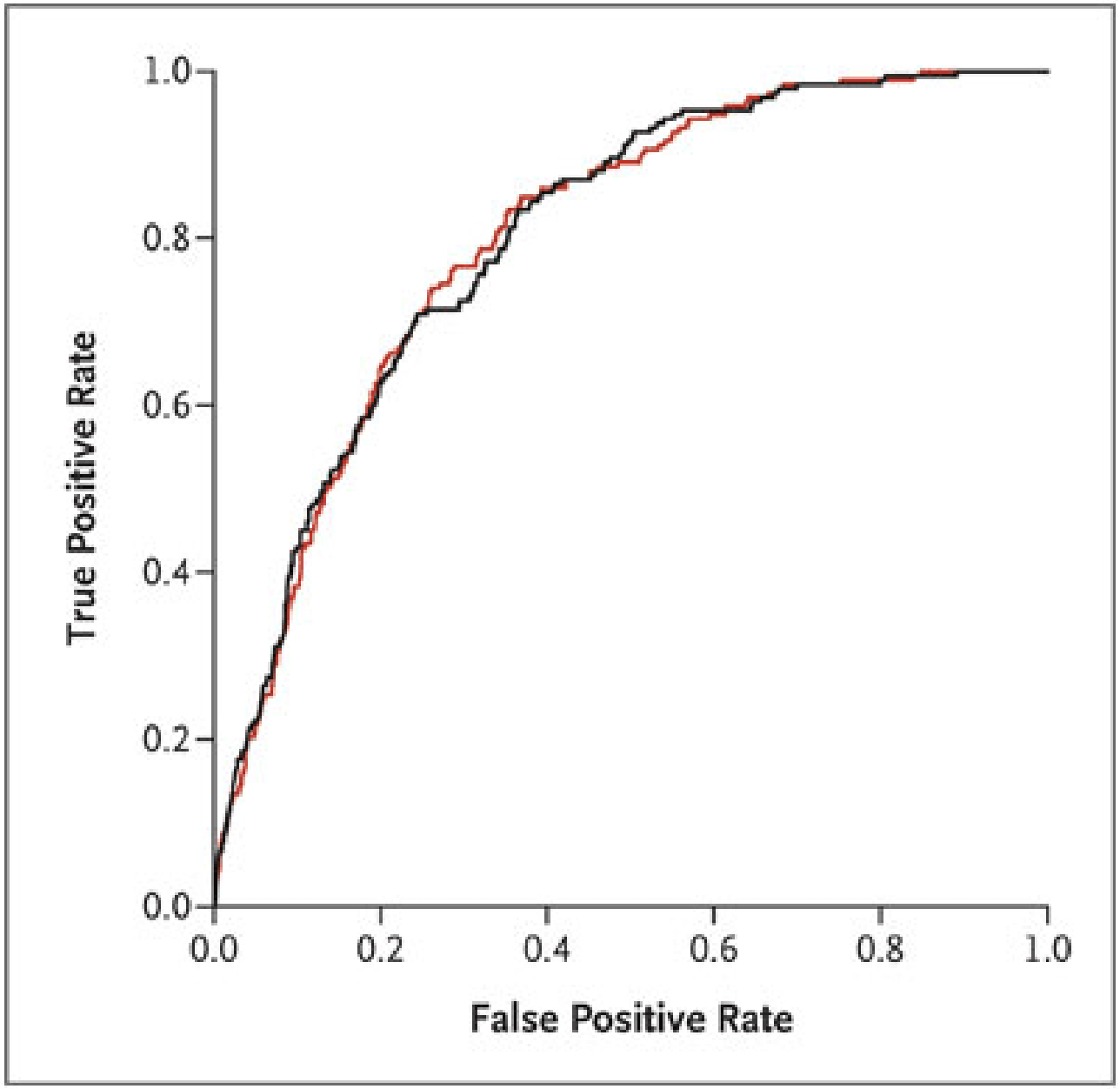
Not unlike forest plot in meta-analysis



The graph is used to identify particular haplotype with strong effect on phenotype



A random colour scheme can be used, highlight or identify points of interests



ROC curves for MI, stroke and death with (black)/without (red) genotype.

Kathiresan et al. *NEJM 2008*

Meta-analysis

- Meta-analysis combines the results of many studies which address a research hypothesis. It is therefore commonly used to increase statistical power as well improve the precision of estimates,
- It is particularly useful for common diseases with many genes each having small effects and has been used in linkage analysis involving both candidate genes and genome scanning, as well as association and GWAS.
- In meta-analysis, the study-specific effect is not a concern (e.g. from logistic as well as Cox regression models with and without adjustment for covariates) unless the estimate and its variance are available. When sample size in each study is fairly large, the Central Limit Theorem ensures that distribution of estimate from each study is approximately normal.
- The key parameter of the meta-analysis model is the variance of the random effect, which reflects the variation between studies. A homogeneity test is available with an extract confidence level, the Q-test.

Simple meta-analysis model

- Suppose a set of studies, i=1,…,n, each provides a value (estimate) of interest yi along with its variance si, a naïve approach to combine them is the average of yi but this do not take into account precision, so a better approach is to take a (inverse variance) weighted average but again that could be sensitive to studies with smaller variances.
- A better meta-analysis model is to treat yi as an observation and the within-study variance as a fixed number, meanwhile allow for a variation (bi) between studies with an known variance, or more specifically. $Yi=\beta+bi+\varepsilon i$. If bi and εi are normally distributed, this is equivalent to $yi \sim N(\beta,\sigma^2+\sigma i^2)$, and β is the common effect, σ^2 is the heterogeneity parameter.
- Following Gauss-Markov theorem, if σ^2 is known, the estimated (inverse variance) weighted average will be BLUE with variance estimator $1/\Sigma(\sigma^2+\sigma i^2)^{-1}$, assuming the σ^2 is known. When $\sigma^2=0$, we have WLS estimator above, and when σ^2=infinity, we have the simple average.

When σ^2 is unknown

- The meta-analysis has a log-likelihood with no analytical solution,

$$l(\beta, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n}\left[\ln(\sigma^2 + \sigma_i^2) + \frac{(y_i - \beta)^2}{\sigma^2 + \sigma_i^2}\right]$$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n}\frac{y_i - \beta}{\sigma^2 + \sigma_i^2}, \quad \frac{\partial l}{\partial \sigma^2} = -\frac{1}{2}\sum_{i=1}^{n}\left[\frac{1}{\sigma^2 + \sigma_i^2} - \frac{(y_i - \beta)^2}{(\sigma^2 + \sigma_i^2)^2}\right],$$

$$\frac{\partial^2 l}{\partial \beta^2} = -\sum_{i=1}^{n}\frac{1}{\sigma^2 + \sigma_i^2}, \quad \frac{\partial^2 l}{\partial \beta \partial \sigma^2} = -\sum_{i=1}^{n}\frac{y_i - \beta}{(\sigma^2 + \sigma_i^2)^2},$$

$$\frac{\partial^2 l}{\partial \sigma^4} = \frac{1}{2}\sum_{i=1}^{n}\left[\frac{1}{(\sigma^2 + \sigma_i^2)^2} - \frac{2(y_i - \beta)^2}{(\sigma^2 + \sigma_i^2)^3}\right].$$

Problem with MLE

- The (2,2)th element of Hessian matrix my be positive, so that the Newton-Raphson algorithm may fail. It may also be that the log-likelihood function has several local maxima.

$$\mathbf{H} = -\begin{bmatrix} \sum_{i=1}^{n} \frac{1}{\sigma^2+\sigma_i^2} & \sum_{i=1}^{n} \frac{y_i-\beta}{(\sigma^2+\sigma_i^2)^2} \\ \sum_{i=1}^{n} \frac{y_i-\beta}{(\sigma^2+\sigma_i^2)^2} & \frac{1}{2}\sum_{i=1}^{n}\left[\frac{2(y_i-\beta)^2}{(\sigma^2+\sigma_i^2)^3} - \frac{1}{(\sigma^2+\sigma_i^2)^2}\right] \end{bmatrix}$$

- The negative expected Hessian matrix of the log-likelihood function, the information matrix is always positive definite, so that Fisher scoring algorithm is more reliable.

$$\mathcal{I} = -E(\mathbf{H}) = \begin{bmatrix} \sum_{i=1}^{n} \frac{1}{\sigma^2+\sigma_i^2} & 0 \\ 0 & \frac{1}{2}\sum_{i=1}^{n} \frac{1}{(\sigma^2+\sigma_i^2)^2} \end{bmatrix}$$

Fisher score algorithm

- The block-diagonal form of the information matrix suggests that we can maximize log-likelihood function separately for β and σ^2

$$\widehat{\beta}_{s+1} = \left(\sum_{i=1}^{n} \frac{1}{\widehat{\sigma}_s^2 + \sigma_i^2} \right)^{-1} \sum_{i=1}^{n} \left(\frac{\widehat{\beta}_s}{\widehat{\sigma}_s^2 + \sigma_i^2} + \frac{y_i - \widehat{\beta}_s}{\widehat{\sigma}_s^2 + \sigma_i^2} \right)$$

$$= \left(\sum_{i=1}^{n} \frac{1}{\widehat{\sigma}_s^2 + \sigma_i^2} \right)^{-1} \sum_{i=1}^{n} \frac{y_i}{\widehat{\sigma}_s^2 + \sigma_i^2},$$

$$\widehat{\sigma}_{s+1}^2$$

$$= \left(\sum_{i=1}^{n} \frac{1}{(\widehat{\sigma}_s^2 + \sigma_i^2)^2} \right)^{-1} \sum_{i=1}^{n} \left[\frac{\widehat{\sigma}_s^2}{(\widehat{\sigma}_s^2 + \sigma_i^2)^2} + \frac{(y_i - \widehat{\beta}_s)^2}{(\widehat{\sigma}_s^2 + \sigma_i^2)^2} - \frac{1}{\widehat{\sigma}_s^2 + \sigma_i^2} \right]$$

$$= \left(\sum_{i=1}^{n} \frac{1}{(\widehat{\sigma}_s^2 + \sigma_i^2)^2} \right)^{-1} \sum_{i=1}^{n} \left[\frac{(y_i - \widehat{\beta}_s)^2 - \sigma_i^2}{(\widehat{\sigma}_s^2 + \sigma_i^2)^2} \right]$$

Restricted maximum likelihood (REML)

- The log-likelihood function has the following form,

$$l_R(\beta, \sigma^2) = -\frac{1}{2}\left\{\sum_{i=1}^{n}\left[\ln(\sigma^2+\sigma_i^2)+\frac{(y_i-\beta)^2}{\sigma^2+\sigma_i^2}\right]+\ln\sum_{i=1}^{n}(\sigma^2+\sigma_i^2)^{-1}\right\}$$

- When σ_i^2=constant (balanced model), we have $\widehat{\beta}=\overline{y}$ and

$$\widehat{\sigma}^2=\frac{1}{n-1}\sum(y_i-\overline{y})^2-\sigma_1^2$$

- However, for unbalanced model, REML is biased.
- Since the number of studies is usually small, we prefer REML over ML estimation.

Quadratic unbiased estimation for σ^2

- It has the attractive property of distribution free as appropriate for small number of studies. We start with vector form model,

$$\mathbf{y} = \beta\mathbf{1} + \boldsymbol{\eta}, \quad E(\boldsymbol{\eta}) = \mathbf{0}, \quad cov(\boldsymbol{\eta}) = \sigma^2\mathbf{I} + \boldsymbol{\Lambda} = \mathbf{V},$$

$$\text{where } \boldsymbol{\Lambda} = diag(\sigma_1^2, ..., \sigma_n^2) \text{ and } \mathbf{1} = (1, ..., 1)'$$

- Start with $\Sigma_{i=1}^{n}(y_i - \bar{y})^2$ and equate to its expected value, we have the unweighted unbiased estimator of the random effect variance, a weighted version is DerSimonian and Laird (1986)

$$\widehat{\sigma}^2_{UMM} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 - \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$$

$$\widehat{\sigma}^2_{WMM} = \frac{\sum_{i=1}^{n}\sigma_i^{-2}(y_i - \widehat{\beta}_0)^2 - (n-1)}{\sum_{i=1}^{n}\sigma_i^{-2} - \frac{\sum_{i=1}^{n}\sigma_i^{-4}}{\sum_{i=1}^{n}\sigma_i^{-2}}}$$

Comparison of estimators for meta-analysis

- The Mininum Norm Quadratic Unbiased Estimation (MINQUE) is to seek a σ^2 in the following form but it turns out to be unweighted moment estimator.

$$\widehat{\sigma}^2 = \mathbf{y}'\mathbf{A}\mathbf{y} - c$$

- Although UMM is better when studies vary significantly, WMM is preferable for minor heterogeneity as is the case with most meta-analysis.
- In the case of σ^2 in the range (0,1), MLE is generally negatively biased, whereas REML and WMM estimators are very close. For this reason WMM is preferable in general.

Statistical inference

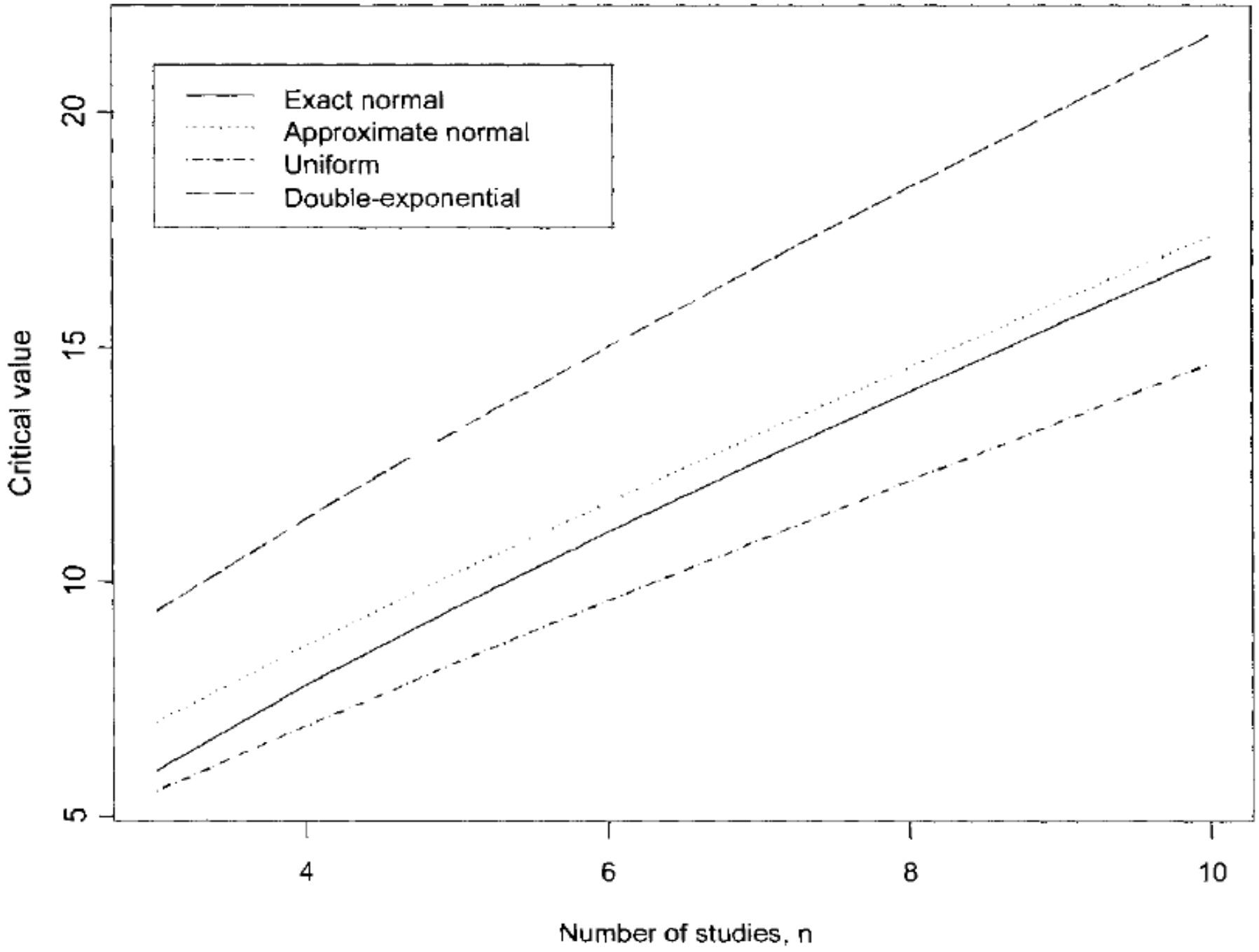
- This is with respect to two major statistical questions: are studies homogeneous ($\sigma^2=0$)? Is the genetic effect statistically significant ($\beta=0$)?
- For normally distributed ε_i, a test for $\sigma^2=0$ with exact significant level suggested by DerSimonian and Laird (1986),

$$Q = \sum_{i=1}^{n} \sigma_i^{-2}(y_i - \widehat{\beta}_0)^2 \sim \chi^2(n-1)$$

- A test based on kurtosis in the case of nonnormal data such that

$$Q \simeq \mathcal{N}(n, (\kappa - 1)n)$$

- The kurtosis for symmetric distribution characterizes how sharp the density is around zero: the sharper, the heavier the tails. $k(X)=E(X^4)/V^2(X)$, e.g., 1.8 (uniform), 3 (normal), 6 (double-exponential).


Exact normal
Approximate normal
Uniform
Double-exponential
Critical value
5
10
15
20
4
6
8
10
Number of studies, n

Hypothesis testing for common effect β

- This is achieved with Wald test

$$Z = \frac{\widehat{\beta}}{\sqrt{v}} = \frac{\sum y_i(\widehat{\sigma}^2 + \sigma_i^2)^{-1}}{\sqrt{\sum(\widehat{\sigma}^2 + \sigma_i^2)^{-1}}} \simeq \mathcal{N}(0,1)$$

- An approximate 95%CI is constructed similarly, but it has been shown that confidence intervals based on the profile-likelihood

$$l_p(\beta) = -\frac{1}{2}\sum_{i=1}^{n}\left[\ln(\sigma^2(\beta) + \sigma_i^2) + \frac{(y_i - \beta)^2}{\sigma^2(\beta) + \sigma_i^2}\right]$$

- to solve $l_p(\beta) = l_{\max} - 0.5\chi^2_{1-\alpha}(1)$ is better.

A measure of heterogeneity

- I^2 statistic is a measure for between-study heterogeneity, i.e., the percentage of total variation across studies due to heterogeneity rather than chance. It can be more informative than the commonly-used Cochran Q test, which has limited power when the number of studies is small.
- I^2 is calculated as: $I^2 = 100\% \times (Q-df)/Q$, where Q is Cochran's heterogeneity statistic and df, the degrees of freedom. Negative values of I^2 are put equal to zero so that it lies between 0% and 100%. Note that df is the mean of a Chi-square distribution and therefore that of Q if the distribution assumption is valid (Higgins et al. *BMJ* 327, 557).

Meta-analysis (fixed effects)

```
data test;
        input studyid lor est;
        col=_n_; row=_n_;
        value=est;
Cards;
…  data for 15 studies …
run;
proc mixed method = ml data=test;
         class studyid;
         model lor = / s cl;
         repeated / group = studyid;
         parms / parmsdata=test eqcons=1 to 15;
run;
```

Meta-analysis (random effects)

```
proc mixed data=test covtest; /*no specification of 15*/
        class studyid;
        model lor = / s cl outp=predp outpm=predm;
        repeated diag / r;
        random studyid / g gdata = test s v;
        ods output CovParms=cp G=G R=R V=V
                        SolutionF=SF SolutionR=SR;
run;
data predp;
      set predp; pvalue=probnorm(resid/stderrpred);
run;
data predm;
      set predm;pvalue=probnorm(resid/stderrpred);
run;
```

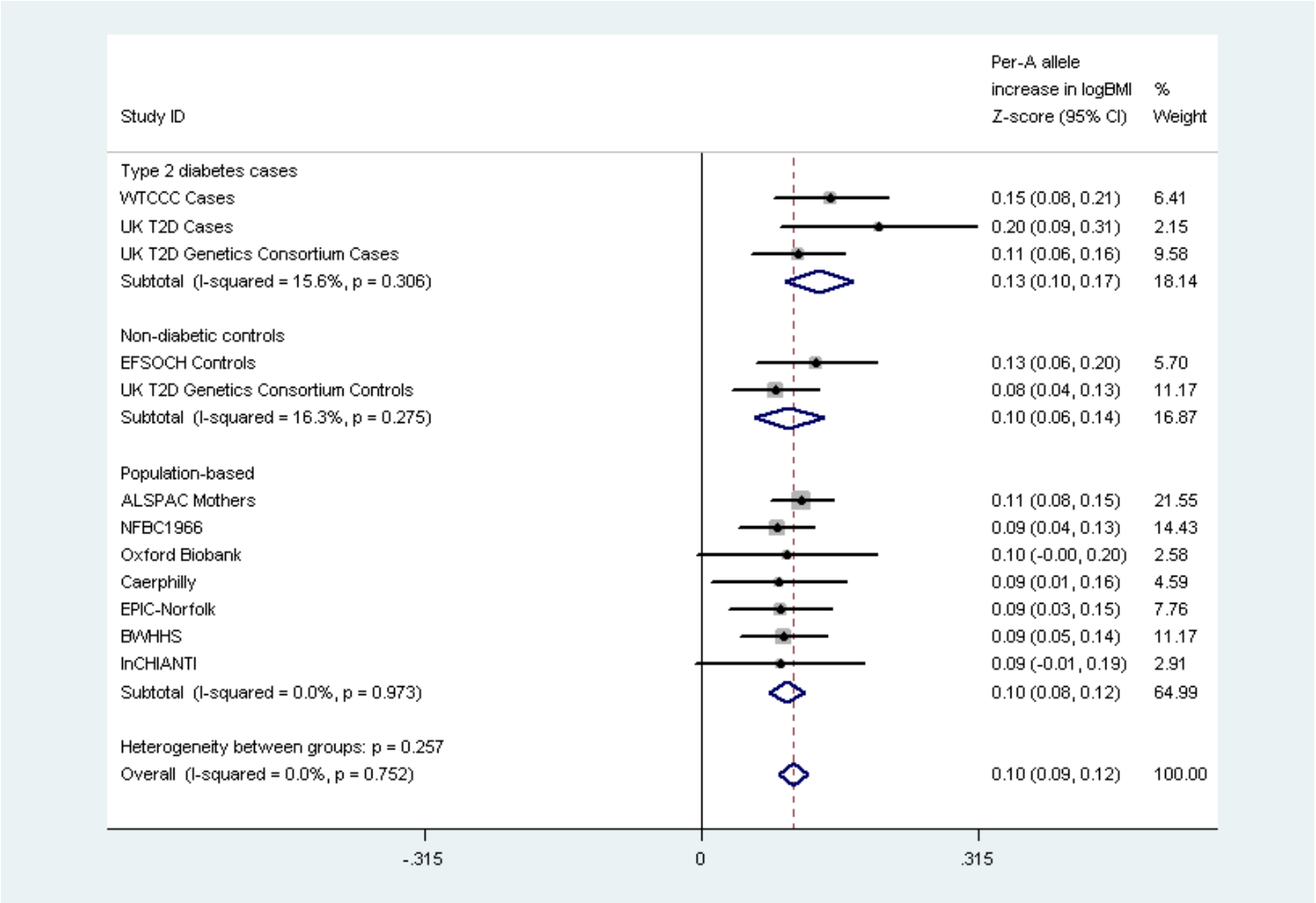
Case-studies -- FTO, BMI and type-2 diabetes

In the Wellcome Trust Case Control Consortium study of 1924 U.K. type 2 diabetes patients and 2938 population controls for 490,032 autosomal SNPs in the FTO (fat mass and obesity associated) gene region on chromosome 16 were strongly associated with T2D (rs9939609, OR = 1.27; 95% CI = 1.16 to 1.37; P = 5 × 10−8).

This association was replicated in a further 3757 T2D cases and 5346 controls (OR = 1.15; 95% CI = 1.09 to 1.23; P = 9 × 10−6). The diabetes-risk alleles at FTO were strongly associated with increased BMI. In the replication samples, the association between FTO SNPs and T2D was abolished by adjustment for BMI (OR = 1.03; 95% CI = 0.96 to 1.10; P = 0.44), which suggests that the association of these SNPs with T2D risk is mediated through BMI.

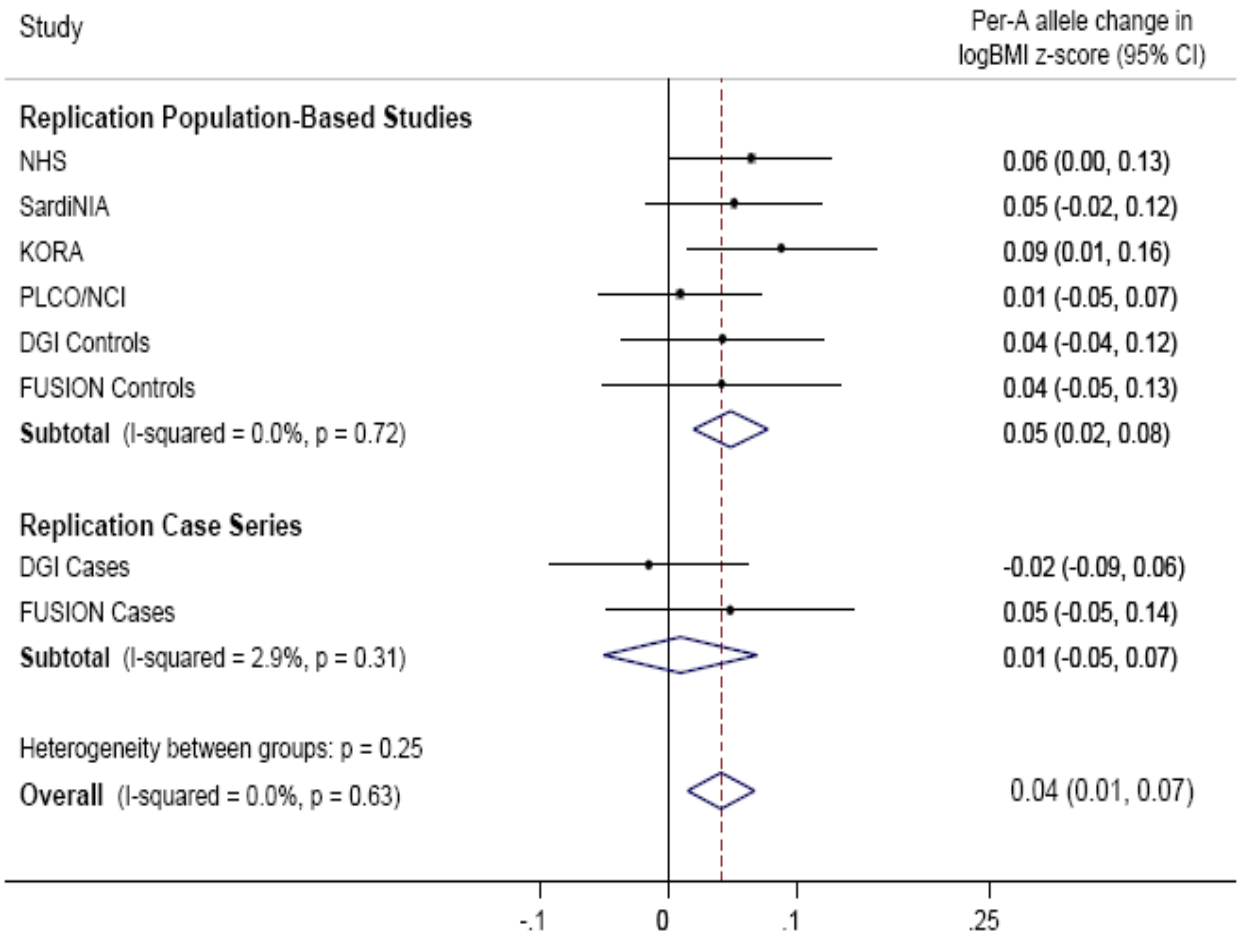
Frayling et al. (2007) *Science* 316: 889-894

Meta-analysis of rs9939609 and BMI



per-A allele effect size on BMI, in log10BMI Z-score units

Meta-analysis of rs17700633 and BMI



Ruth et al. (2008) Nat Genet 40(6):768-775

Per-A allele effect size on BMI in 10 adult population-based studies (n = 13,240) and 3 case series (n = 2,638) in log10BMI Z-score units.